28 April 2020

# DQF BI Bulletin #6 - April 2020

## The Impact of Text Complexity on Localization




In his famous book 'Thinking, Fast and Slow', Daniel Kahneman has many examples of how cognitive strain makes our thinking slow. There is only a limited set of concepts that we can work with at the same time, and under mental strain, people become less critical. For example, there is the experiment where someone has to add the number 3 to a series of 4 different digits and must repeat this exercise over and over on the result of the previous addition. On top of that, the calculation has to be made at a certain pace of a few seconds for each calculation. Trying to remember the four digits each time is heavy lifting for the brain. His point is that under such strain people would like to release the burden a bit.

The translation of complex sentences might require the same mental powerlifting from the translator, as he/she juggles with different concepts in a single sentence, and needs to figure out if a machine translation has transferred the concepts correctly from source to target. Let's see if there is something to be said about the text complexity when researching DQF data. For the sake of comparison, we are focusing only on segments that are pre-translated with machine translation, and only on DQF's most frequent language combinations. This includes English to Simplified Chinese, French, German, Italian, Japanese, Korean, Latvian, Lithuanian, Brazilian Portuguese, Russian, and Mexican Spanish.

## The Role of Content Type

The DQF Dashboard has been designed to give you immediate insights into the most important metrics of your localization process. It shows the trends in your workflows over time and you can make quick comparisons. You can see which languages in your production process are doing well, and which ones might need some extra attention. Or you can monitor how your machine translation engine is performing over time. Useful for your day-to-day business, but there is much more richness in the data for those who want to dig deeper. Even though it might seem that the insights are not immediately actionable, they surely tell a story of how localization processes work.

One of the basic distinctions DQF makes in translation projects concerns the content type of the text. It is a useful distinction that becomes clear as soon as we compare the productivity rates of content types. Most of us can guess that translating, for example, marketing material takes much more time than translating user manuals.

DQF data confirms this assumption. If we create a level playing field and only compare translations from the higher quality tier, that started with suggestions from machine translation, we see much lower productivity in projects with a marketing content type. Each edit simply takes more time, and at the end of the day, fewer words are translated.

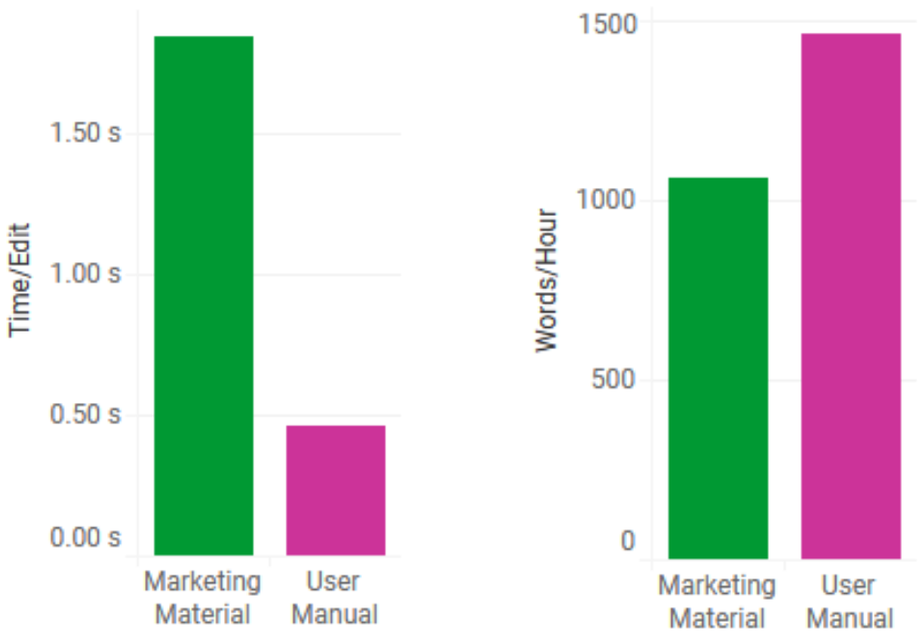



Figure 1: Comparing translation productivity, based on content type, only using MT output

The creativity that is needed to provide a good marketing translation is probably one of the main factors in the lower productivity numbers. However, apart from approaching translations through the DQF categories, would it be possible to look at the real content of the translations? Would it be possible to see, for example, how text complexity impacts productivity?

Complexity of text is hard to measure. It is affected by many qualitative dimensions. Think of the levels of meaning in the text: the structure, clarity, conventionality, and so on. Usually, human judgment is required to give an estimation of these dimensions of complexity.

However, there are also some dimensions of complexity that are easier to quantify. For example, the length of the sentence, and the average length of the words in the text. And this is an interesting ground for research. The DQF database seems to be well-suited for this type of research with the data it collected over the years.

# The Odd Segments...

It seems simple: measure the length of a sentence, and try to figure out the impact of sentence length on the productivity metrics. But let's stop for a moment, and try to think about a 'normal' sentence. You probably think of a group of around ten to twenty words, all made up from the letters of the alphabet, some common punctuation marks, and finally a full stop to conclude the sentence.

But once we start looking at all source segments in real localization projects, there appears to be a big group of odd ones out there. They don't fit this description. Certainly among the very long sentences and the very short sentences, there is a tendency that a large portion of the text is taken by numbers, formatting tags, and characters such as hashtags and underscores. Translators see this type of text a lot as they work through all sorts of specifications, tables, or exports of heavily formatted text.

These atypical text elements can be an extra challenge to work with, depending on the project and the amount, to make sure that in the final translation these elements are kept in the right place. Though this part of localization is largely disregarded in scientific literature, it is a text feature that might add to the complexity, in the sense that it seemingly decreases the clarity of the text.

Let's think of a way to quantify segments on their conformance to what is generally seen as a 'typical text'. On one end of the continuum, there are the sentences that are entirely made up of alphabetical characters, spaces, and some common punctuation, such as commas and full stops. On the other end, there are the sentences that are made up of formatting elements within "<" and ">" brackets, URLs, and as far as the rest of the text is concerned, contain a large portion of numbers and other non-alphabetical characters.

A simple approximation of what I would call the 'prose ratio', would be to calculate what percentage of the total number of characters in a segment is taken by alphabetical characters and common punctuation symbols outside of formatting tags or URLs. This prose ratio is 100% for a typical sentence with only human-readable words, and close to 0% whenever that sentence contains just numbers and tags.

It appears that although most segments in the DQF database can be described as 'typical text', there is a considerable number of segments with a prose ratio lower than 80%, that contain tags and technical data, and not much quickly readable text.

Around 20% of segments in our sample have a prose ratio lower than 80%, which means that text with low prose ratio is actually a substantial part of the translator's job.

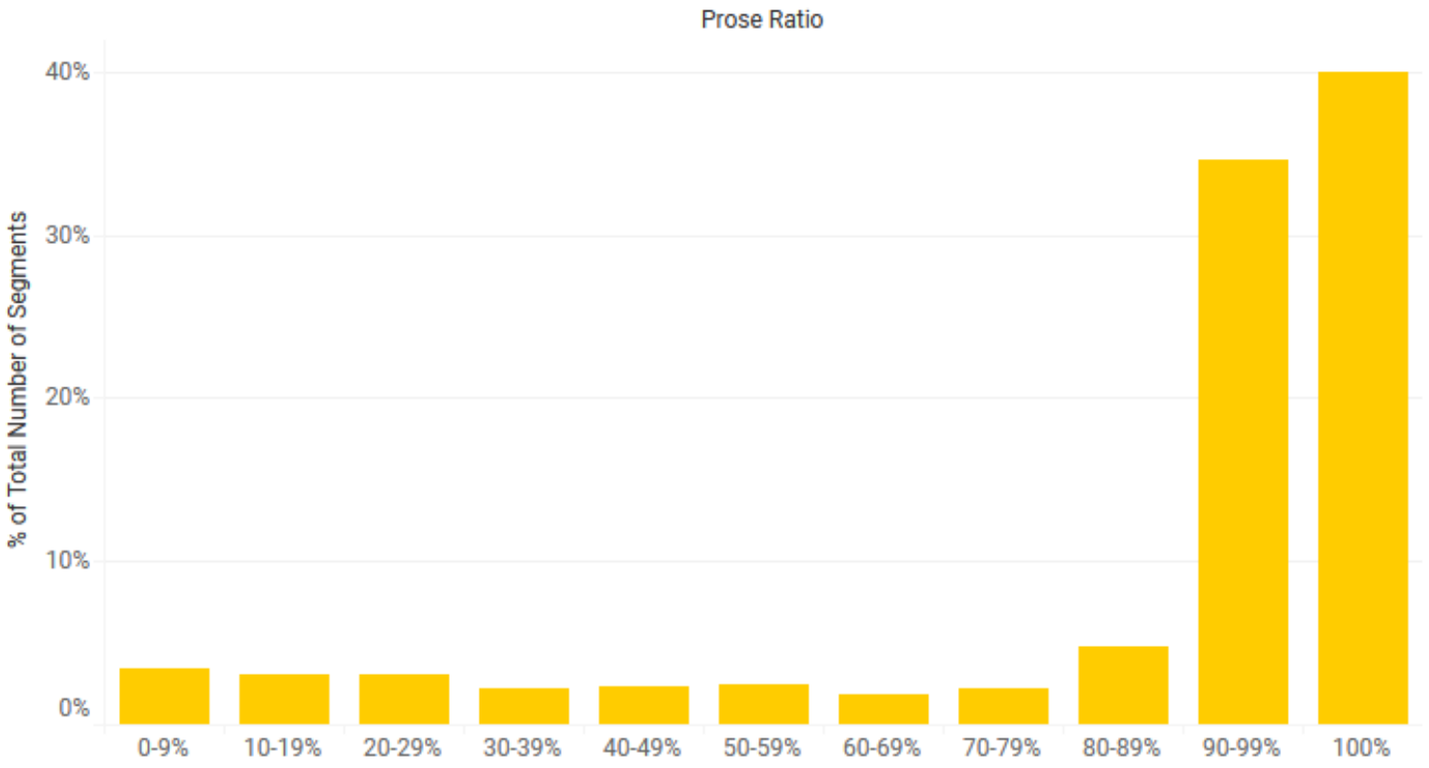



Figure 2: Distribution of segments based on prose ratio

In our selection of segments pre-translated with machine translation, it turns out that prose ratio has a quite strong influence on the edit density. Translators make fewer edits per 1000 characters whenever the prose ratio is low.

The edit density seems to settle whenever the prose ratio gets above 70%. The correlation is quite strong. This seems to show that MT is performing better in text with lower prose ratios, and requires more editing in segments with a high prose ratio.

The same can be said about the editing time, as edits seem to take longer whenever the prose ratio is high. However, this is a much weaker correlation.

The effect of the prose ratio on the productivity rate is very diffuse. If there is any trend, then it seems to suggest, in accordance with the above, that high prose ratios decrease the number of words translated per hour.

Part of this weak correlation might be explained by the observation that segments in the lower prose ratio are very different in nature. They can be extremely long, or short, and they usually contain very few of what can be counted as words, so the “words per hour” as a measurement for this type of segment might not be such a good metric.

## … and the long ones

As mentioned above, segment length is one of the quantifiable parameters of text complexity. We see that segment length has a special relationship with prose ratio. Both extremely short and extremely long segments tend to have a low prose ratio. The average prose ratio can go below an average of 50% for segments with less than 10 words or when segments become extremely long, with more than 40 words.

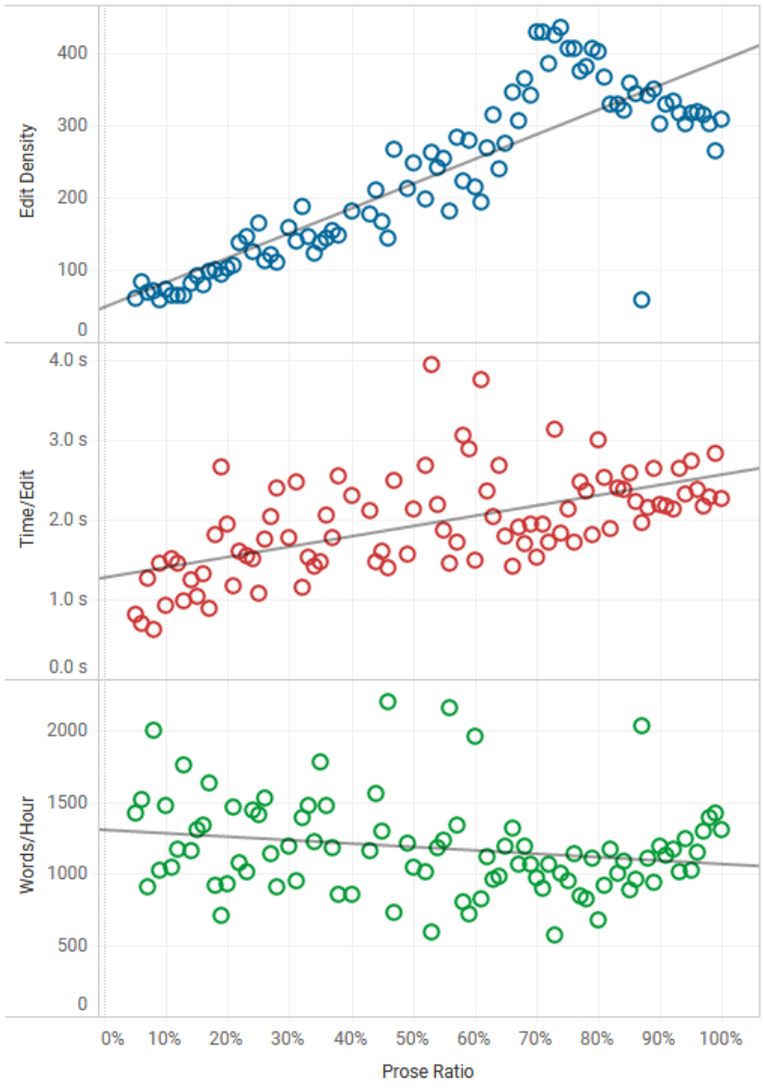



Figure 3: Effect of prose ratio on productivity parameters

If we want to measure the effect of segment length on productivity, it is better to filter out the odd ones such as the segments with a low prose ratio.

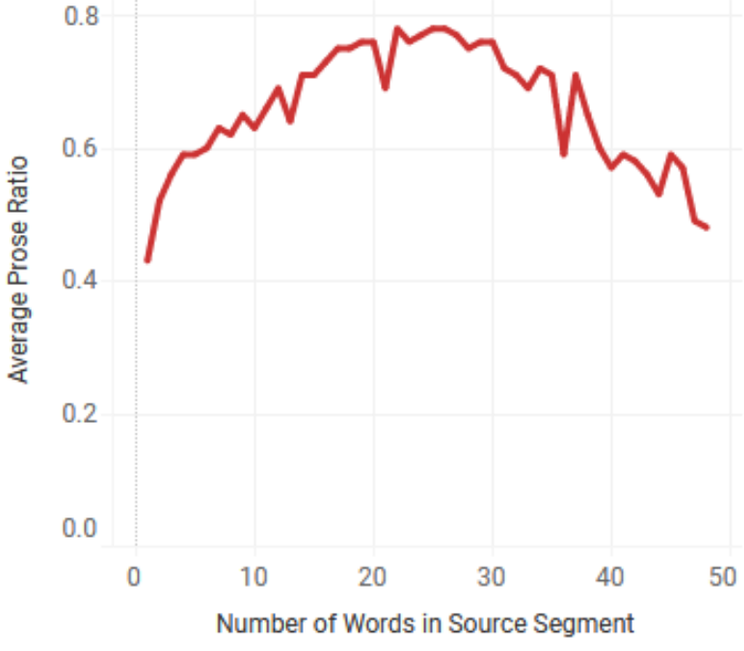



Figure 4: Length of segments and their average prose ratio

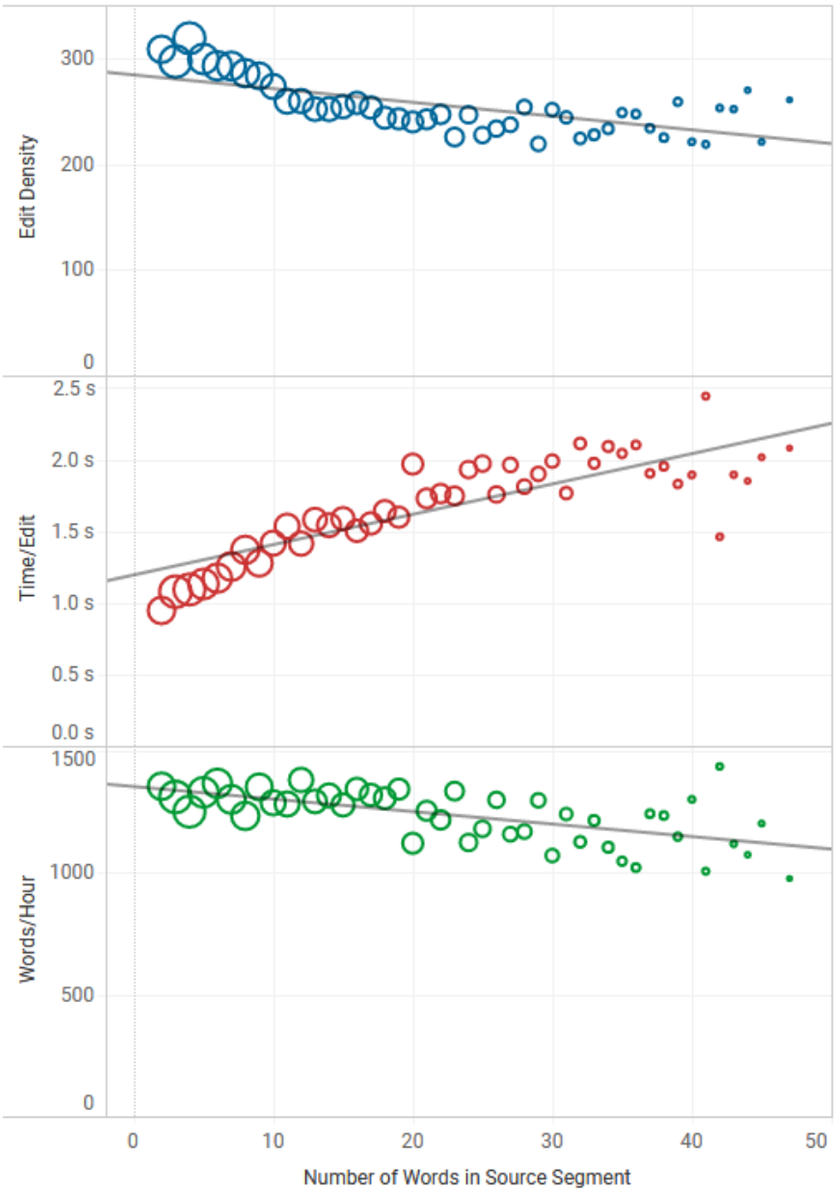



Figure 5: Effect of segment length on productivity parameters; size of the circle indicates the relative frequency of segments of that length

With the lower prose ratio out of the way, the length of the segments has a few remarkable effects on the translation. First of all, longer segments have a lower edit density than shorter ones. It is to be expected that longer sentences translated by MT engines, due to lower performance, would need more post-editing. However, it also shows that in lengthier segments the time spent per edit is much longer. Around 25 words per segment there seems to be a plateau in the charts, that indicates that time spent per edit does not go much above 2.0 seconds, whereas for very short segments, the time spent per edit is close to 1.0 second. This seems to confirm that the length of the segment certainly puts an extra cognitive strain on the translator, and as a consequence slows down the editing process.

With the edit density decreasing as the segment length grows, while the time per edit increasing, the number of words translated in an hour is not so much affected. There is a trend showing that longer sentences do slow down the translator, but the correlation is not as strong.

It would be interesting to see why longer segments, even when filtering out the ones with a low prose ratio, have a lower edit density. So far, we can only speculate. One explanation might be that the effort for edits simply becomes too much. Two seconds for each edit being the upper limit, it might be that in the process of production and tight deadlines, the suggestions from the machine translation are easier accepted. Sifting through long sentences and finding the right translation, which might include complete restructuring of the sentence, might be too much of a strain. Of course to test this hypothesis, there should be more extensive research in the output quality of translations to see if longer segments might need more corrections. But the results of this analysis seem to confirm this.

## Innovation Through Data

Going beyond the directly available data, and doing your own research on the translation data can be an interesting opportunity to look at your workflow in a different way. In this bulletin we often report on the trends in the DQF database. But exploration can go further than that. To dig deeper into the data and show some new correlations can be inspiring and fuels innovation in the longer run.

In this bulletin we analyzed the content of the source segments. Going through the content makes you aware of the patterns that exist in what is basically unstructured data. It takes creativity to make sense of these patterns. In this respect metrics like the prose ratio show us that different types of segment content have a fundamental impact on the nature of the translation process.

Furthermore, it also shows that segment length has a distinct influence on the time that is spent on each edit. Lower edit density in the longer sentences, contrary to what would be expected, might indicate that longer sentences need some more attention, either in extra correction rounds, or else suggest that further MT training is needed. But in either case, the starting point is the data analysis.

